Knowledge Extraction for Clinical Question Answering: Preliminary Results

Dina Demner-Fushman[1,3] **and Jimmy Lin**[2,3]
[1]Department of Computer Science
[2]College of Information Studies
[3]Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20742, USA
demner@cs.umd.edu, jimmylin@umd.edu

Abstract

The combination of recent developments in question answering research and the unparalleled resources developed specifically for automatic semantic processing of text in the medical domain provides a unique opportunity to explore complex question answering in the clinical domain. In this paper, we attempt to operationalize major aspects of evidence-based medicine in the form of knowledge extractors that serve as the fundamental building blocks of a clinical question answering system. Our evaluations demonstrate that domain-specific knowledge can be effectively leveraged to extract PICO frame elements from MEDLINE abstracts. Clinical information systems in support of physicians' decision-making process have the potential to improve the quality of patient care in real-world settings.

Introduction

The focus of question answering research is shifting away from simple fact-based questions that can be answered with relatively little linguistic knowledge to "harder" questions that require reasoning and gathering information from multiple sources. General purpose reasoning on anything other than superficial lexical relations is exceedingly difficult because there is a vast amount of world and commonsense knowledge that must be encoded, either manually or automatically, to overcome the brittleness often associated with long chains of evidence. However, the availability of rich existing knowledge sources and ontologies in certain domains presents an interesting opportunity for question answering systems. How might one go about leveraging these resources effectively?

We explore this research problem in the clinical domain, which is well-suited for experiments with knowledge-based question answering techniques for several reasons. First, understanding of the domain has already been codified in the Unified Medical Language System® (UMLS) (Lindberg, Humphreys, & McCray 1993). Second, software for utilizing this ontology already exists: MetaMap (Aronson 2001) identifies concepts in free text, while SemRep (Rindflesch & Fiszman 2003) extracts relations between the recognized concepts. Both systems utilize and propagate semantic information from UMLS knowledge sources: Metathesaurus®, the Semantic Network, and the SPECIALIST lexicon. The 2004 version of the UMLS Metathesaurus contains information about over 1 million biomedical concepts and 5 million concept names from more than 100 controlled vocabularies. The Semantic Network provides a consistent categorization of all concepts represented in the UMLS Metathesaurus. Third, the framework of evidence-based medicine (Sackett *et al.* 2000) provides a task-based model of the clinical information-seeking process; the PICO frame for capturing well-formulated clinical queries (described later) can serve as the knowledge representation that bridges the needs of clinicians and analytical capabilities of a system. The confluence of these many factors makes clinical question answering a very exciting area of research.

Furthermore, the need to answer questions related to patient care at the point of service has been well studied and documented (Covell, Uman, & Manning 1985; Gorman, Ash, & Wykoff 1994). According to Ely *et al.* (2005), the desirable features of a system capable of providing answers to clinical questions are:

- "comprehensive resources that answer questions likely to occur in practice with emphasis on treatment and bottom-line advice", and
- the ability to "locate information quickly by using lists, tables, bolded subheadings, and algorithms, and by avoiding lengthy, uninterrupted prose".

The MEDLINE® database is ideally suited for addressing the first requirement and is indeed often used by clinicians in that capacity (DeGroote & Dorsch 2003). However, studies have also shown that existing systems for searching MEDLINE are often inadequate and unable to supply clinically-relevant answers in a timely manner (Gorman, Ash, & Wykoff 1994; Chambliss & Conley 1996). Reflecting on Ely *et al.*'s second requirement, it is clear that traditional document retrieval technology applied to MEDLINE abstracts is insufficient for satisfactory information access; research and experience point to the need for systems that automatically analyze text and return only the relevant information, appropriately summarizing and fusing segments from multiple texts. Such a system should also rank results based on their relevance to the clinical task, taking into ac-

count such factors as the quality of research results and recency of the article. In short, clinicians would greatly benefit from advanced question answering capabilities to provide decision support in the patient care process.

This paper reports ongoing efforts to develop and deploy clinical question answering systems, which build on previous related projects we have worked on (Demner-Fushman *et al.* 2004). We focus our attention here on operationalizing the process of evidence-based medicine in terms of knowledge extractors, which serve as the building blocks of an end-to-end clinical question answering system. More specifically, this paper describes techniques for extracting population, problem, intervention, comparison, and outcome from MEDLINE abstracts. These elements, combined with metadata already associated with MEDLINE citations, determine the relevance of a particular abstract with respect to clinicians' questions.

With respect to the broader research question concerning the role of knowledge in question answering, we demonstrate that simple, appropriate uses of domain knowledge can simplify the task of extracting relevant semantic information from text. Evaluations show that our knowledge extraction techniques achieve respectable performance and serve as a solid foundation for future work.

Evidence-Based Medicine

Evidence-based medicine (EBM) is a widely-accepted paradigm for medical practice that stresses the importance of evidence from patient-centered clinical research in the health care process. Clinical evidence provides the information necessary to develop physicians' individual expertise, which in turn results in higher quality patient care. We seek to develop decision-support systems that complement this paradigm of medical practice.

Evidence-based medicine offers three orthogonal views of the domain that, when taken together, provide a framework for codifying the knowledge involved in answering questions related to patient care. These three complementary views are outlined below.

The first view describes the four main clinical tasks that physicians engage in: therapy, diagnosis, etiology, and prognosis. Terms and the types of studies relevant to each of the four tasks have been extensively studied by the Hedges Project at the McMaster University (Wilczynski, McKibbon, & Haynes 2001). The results of this research are implemented in the PubMed Clinical Queries tools, which can be used to retrieve task-specific citations.

The second view is independent of the clinical task and pertains to the structure of a well-built clinical question. The following four components have been identified as the key elements of a question related to patient care (Richardson *et al.* 1995):

1. What is the primary problem or disease? What are the characteristics of the patient (e.g., age, gender, or co-existing conditions)?
2. What is the main intervention (e.g., a diagnostic test, medication, or therapeutic procedure)?
3. What is the main intervention compared to (e.g., no intervention, another drug, another therapeutic procedure, or a placebo)?
4. What is the effect of the intervention? Were the patient's symptoms relieved or eliminated? Side effects reduced? Cost reduced? etc.

These four elements are often referenced with a mnemonic PICO, which stands for Patient, Intervention, Comparison, and Outcome.

Finally, the third view serves as a tool for appraising the strength of evidence presented in the clinical study, i.e., how much confidence should a physician have in the results? Several taxonomies for appraising the strength of evidence based on the type and quality of the study have been developed. We chose the Strength of Recommendations Taxonomy (SORT) as the basis for determining the potential upper bound on the quality of evidence, due to its emphasis on the use of patient-oriented outcomes and its attempt to unify other existing taxonomies (Ebell *et al.* 2004). There are three levels of recommendations according to SORT:

1. A-level evidence is based on consistent, good quality patient outcome-oriented evidence presented in systematic reviews, randomized controlled clinical trials, cohort studies, and meta-analysis.
2. B-level evidence is inconsistent, limited quality patient oriented evidence on the same types of studies
3. C-level evidence is based on disease-oriented evidence or studies less rigorous than randomized controlled clinical trials, cohort studies, systematic reviews and meta-analysis.

Any question answering system designed to support the practice of evidence-based medicine must be sensitive to the multifaceted considerations that go into evaluating an abstract's relevance to a clinical query. As a component of a clinical question answering system, we have developed knowledge extractors that identify PICO frame elements from MEDLINE abstracts and classify their evidence grade level (corresponding to the second and third views of evidence-based medicine). We first describe these techniques and then present our evaluation results.

Extracting PICO Frame Elements

Evidence-based medicine provides a pre-existing domain model for encoding the semantic knowledge necessary to answer a clinical question. In particular, the PICO frame describes the structure of a well-built clinical query, and can serve as the core organizing knowledge structure of a question answering system: the information seeking process can be viewed as semantic unification between partially instantiated PICO query frames and corresponding frames automatically extracted from MEDLINE abstracts. Although clinicians are most often interested in the outcome (of a treatment, for example), the other frame elements are critical in assessing the relevance of a particular study. Thus, the automatic extraction of population, problem, intervention, comparison, and outcome represents a key capability integral to

clinical question answering. This section details extraction modules that identify each of these elements (see example of a completely annotated abstract in Figure 1).

As previously mentioned, software already exists for identifying concepts (MetaMap) and relations (SemRep) in medical texts, which we extensively use in our knowledge extractors. Furthermore, we take advantage of coarser-grained semantic types, Semantic Groups (McCray, Burgun, & Bodenreider 2001), to capture higher-level generalizations. An additional feature we take advantage of (when present) is explicit discourse markers present in some abstracts. These so called structured abstracts were recommended by the Ad Hoc Working Group for Critical Appraisal of the Medical Literature (1987) to help humans assess the reliability and content of a publication and to facilitate the indexing and retrieval processes. These abstracts loosely adhere to the introduction, methods, results, and conclusions format, and summarize a study using sections with the above headings. Although many core clinical journals require structured abstracts, there is a great deal of variation in the actual headings. Even when present, the headings are usually not organized in a manner focused on patient care. In addition, abstracts of many high-quality research remain unstructured. For these reasons, explicit discourse markers are not entirely reliable indicators for the various semantic elements we seek to extract, but must be considered along with other sources of evidence.

The extraction of each PICO frame element relies to a different extent on an annotated corpus of MEDLINE abstracts. The first author of this paper lead an effort in the creation of such a collection at the National Library of Medicine. As will be described below, the population, problem, and the intervention/comparison extraction modules are based on manually constructed rules; the outcome extraction module, in contrast, employs supervised machine learning techniques. These two very different approaches are caused by differences in the nature of the frame elements: whereas problem and intervention can be directly mapped to concepts, and population easily maps to patterns that include concepts, outcome statements are always passages ranging in size from a single clause to eight sentences. The initial goal of our annotation effort was to identify outcome statements in abstract text (Demner-Fushman *et al.*, in preparation). A physician, two nurse practitioners, and an engineering researcher manually identified sentences that describe outcomes within 633 MEDLINE abstracts. The abstracts were retrieved using PubMed and attempted to model different user behaviors ranging from naïve to expert. With the exception of 50 articles retrieved to answer a childhood immunization question, the rest of the articles were retrieved using a disease, for example, diabetes. When emulating an expert user, advanced search features were employed. Of the 633 citations, one hundred abstracts were also fully annotated with population, problem, intervention, and comparison. These one hundred abstracts were set aside as a held-out test set. Of the remaining citations, 275 were used for training and rule derivation, as described in the following sections.

The Two Ps

The PICO framework makes no distinction between the population and the problem, which is rooted in the concept of the population in clinical studies, e.g., the following sentence in a structured abstract: "POPULATION: Fifty-five postmenopausal women with a urodynamic diagnosis of genuine urinary stress incontinence." Although this clause simultaneously describes the population (of which any particular patient can be viewed as a sample therefrom) and the problem, we chose to separate the extraction of the two elements because they are not always described together. Furthermore, many clinical questions ask about a particular problem without specifying a population.

Population Extractor

Population statements are identified using a series of manually crafted rules, based on several assumptions:

- The concept involved in the description of population belongs to the semantic type GROUP or any of its children. In addition, certain nouns are often used to describe study participants in medical texts; for example, an often observed pattern is "subjects" or "cases" followed by a concept from the semantic group DISORDER.
- The number of subjects that participated in the study precedes or follows the concept identified as a GROUP.
- The confidence that a clause with an identified number and GROUP contains information about the population is inversely proportional to the distance between the two entities.
- The confidence that a clause contains the population is influenced by the position of the clause (with respect to headings in the case of structured abstracts and with respect to the beginning of the abstract in the case of unstructured abstracts).

Given these assumptions, the population extractor searches for the following patterns:

- GROUP (n=*number*) (for example, "in 5-6-year-old French children (n=234), Subjects (n = 54)")
- *number** GROUP (for example, "forty-nine infants")
- *number** DISORDER* GROUP? (for example, 44 HIV-infected children)

The confidence of a particular pattern match is a function of both its position in the abstract and its position in the clause from which it was extracted. If a number is followed by a measure, for example, *year* or *percent*, the number is discarded, and pattern matching continues. After the entire document is processed in this manner, the matched pattern with the highest confidence value is retained as the population description.

Problem Extractor

The problem extractor relies on recognition of concepts belonging to the semantic group DISORDER (McCray, Burgun, & Bodenreider 2001). In short, it simply returns a partially ranked list of all concepts recognized as DISORDER. We

evaluate the performance of this simple heuristic on segments of the abstract varying in length: the abstract title only, abstract title and first two sentences, and entire abstract text. Concepts identified in the title are given preference in the ranked ordering of problems, in an effort to distinguish the primary problem from co-occurring conditions.

Intervention/Comparison Extractor

The intervention and comparison frame elements do not require separate processing because they usually belong to the same semantic type. Our intervention extractor simply produces an unordered list of interventions under study, of which one is the main intervention and the rest are comparisons. For convenience, we simply refer to this module as the intervention extractor.

For each of the clinical tasks, semantic relations defined in the UMLS Semantic Network provide strong cues for the intervention: *treats* and *carries out* for THERAPY; *diagnoses* for diagnosis; *causes* and *result of* for etiology; and *prevents* for prognosis. Restrictions on the semantic type of concepts participating in these relations serve as the basis for the intervention extractor rules. For example, the relation THERAPEUTIC OR PREVENTIVE PROCEDURE *treats* DISEASE OR SYNDROME identifies the THERAPEUTIC OR PREVENTIVE PROCEDURE as the intervention for the clinical task therapy. At present, our intervention extractor recognizes concepts belonging to nine semantic types, for example, DIAGNOSTIC PROCEDURE, CLINICAL DRUG, and HEALTH CARE ACTIVITY.

In addition to the semantic type information, the intervention extraction rules take into account positional information. With structured abstracts, the titles, aims, and methods sections are most likely to provide information about the intervention. With unstructured abstracts, relations from the first third of the abstract are heavily favored. Finally, the intervention extractor takes into account the presence of certain cue phrases that describe the aim and/or methods of the study, such as "This * study examines" or "This paper describes". As in the previous modules, information from these different sources are combined using an ad-hoc weighting scheme.

Outcome Extractor

In contrast with the other modules, we approach outcome extraction as a classification task at the sentence level, i.e., for each sentence in an abstract, the outcome extractor predicts whether it states an outcome or not. Our preliminary explorations have lead to a strategy based on an ensemble of classifiers, which include: a rule-based classifier, a unigram "bag of words" classifier, a *n*-gram classifier, a position classifier, a document length classifier, and a semantic classifier. With the exception of the rule-based classifier, all classifiers were trained on the 275 citations from the annotated collection described above.

Knowledge for the rule-based classifier was hand-coded by a registered nurse with 20 years of clinical experience prior to the annotation effort. This classifier outputs a binary decision based on cue phrases such as "significantly greater", "well tolerated", and "adverse events".

The unigram "bag of words" classifier is a Naïve Bayes classifier implemented with the API provided by the MALLET toolkit[1]. This classifier outputs the probability of a class assignment.

The *n*-gram based classifier is also a Naïve Bayes classifier, but it operates on a different set of features. We first identified the most informative unigrams and bigrams using the information gain measure (Yang & Pedersen 1997), and then selected only the positive outcome predictors using odds ratio (Mladenic & Grobelnik 1999). Topic-specific terms, such as rheumatoid arthritis, were then removed. Finally, the list of features was revised by the registered nurse who participated in the annotation effort. This classifier also outputs the probability of a class assignment.

The position classifier returns the maximum likelihood estimate that a sentence is an outcome based on its position in the abstract (for structured abstracts, with respect to the results or conclusions sections; for unstructured abstracts, with respect to the end of the abstract).

The document length classifier returns a smoothed (add one smoothing) probability that a document of a given length (in the number of sentences) contains an outcome statement. For example, the probability that a four sentence-long document contains an outcome statement is 0.25, and the probability of finding an outcome in a ten sentence-long abstract is 0.92. Interestingly, the average length of documents with and without outcome statements differs: the average length of the former is 11.7 sentences, whereas the length of the latter is 7.95 sentences.

The semantic classifier assigns to a sentence an ad-hoc score based on the presence of UMLS concepts belonging to semantic groups highly associated with outcomes such as THERAPEUTIC PROCEDURE or PHARMACOLOGICAL SUBSTANCE. The score is given an ad-hoc boost if the concept has already been identified by MetaMap elsewhere in the abstract (for example, if the problem or intervention were observed in the sentence under consideration).

The output of our basic classifiers are combined using linear interpolation with ad-hoc weights assigned based on intuition. We recognize that our outcome extractor employs a "kitchen sink" approach, but note that this module is mostly the outgrowth of an exploration process in the uncharted solution space for clinical question answering. A more principled approach to outcome extraction will be reserved for future work.

Determining the Strength of Evidence

The potential highest level of the strength of evidence for a given citation can be identified using the publication type and/or MeSH headings pertaining to the type of the clinical study assigned to the article during the indexing process. Table 1 shows our mapping from publication type and MeSH heading to the evidence grade, based on principles defined in the Strength of Recommendations Taxonomy.

Additional information necessary for the final determination of relevance includes the number of study participants, statistical methods involved, randomization, blinding, and

[1] `http://mallet.cs.umass.edu`

Strength of Evidence	Publication Type / MeSH
Level A(1)	Meta-Analysis, Randomized Controlled Trials, Cohort Study, Follow-up Study
Level B(2)	Case-Control Study, Case Series
Level C(3)	Journal Article, Case Report

Table 1: Strength of evidence categories based on Publication Type and MeSH headings.

	correct	unknown	wrong
baseline	53.3%	-	46.7%
extractor	80%	10%	10%

Table 2: Evaluation of the population extractor

	correct	unknown	wrong
abstract title	85%	10%	5%
title + 1st two sentences	90%	5%	5%
entire abstract	86%	2%	12%

Table 3: Evaluation of the problem extractor

the quality of follow-up (Ebell *et al.* 2004). Out of these, the number of subjects is extracted along with the population information. Identification of the statistical, blinding, and randomization methods will be addressed in future work.

Results

This section describes evaluations conducted on the extraction modules. Results are reported in terms of the percentage of correctly identified instances, percentage of instances for which the extractor had no answer, and percentage of incorrectly identified instances. The baselines and gold standards for each extraction module varies, and will be described individually.

Population Extraction

Ninety of the one hundred fully-annotated articles in our collection were agreed upon by the annotators as being clinical in nature, and were used as test data for our population extractor. Since these abstracts were not examined in the rule-creation process, they can be viewed as a blind held-out test set. The output of our population extractor was judged to be correct if it occurred in the same sentence that was annotated as containing the population in the gold standard.

For comparison, our baseline simply returned the first three sentences of the abstract. We considered the baseline correct if any one of the sentences were annotated as containing the population in the gold standard. This baseline was motivated by the observation that the aim and methods sections of structured abstracts are likely to contain the population information. Generally, these sections can be found in the first three sentences of both structured and unstructured abstracts.

The performance of our population extractor is shown in Table 2; note that the evaluation of the baseline is much more lenient than the evaluation of our population extractor.

There were several sources of incorrect and missed populations:

- Not all population descriptions contain a number explicitly, e.g., "The medical charts of all patients who were treated with etanercept for back or neck pain at a single private medical clinic in 2003".
- Not all study populations are population groups, as for example in "All Primary Care Trusts in England."
- Part of speech tagging and chunking errors propagate to the semantic type assignment level and affect the quality of MetaMap output.

Problem Extraction

The goal of the problem extractor is to identify the main problem that calls for the interventions outlined in the abstract. At present, we assume that the main problem is always a DISORDER. Based on this assumption, the gold standard for the problem extractor can be defined using the MeSH headings assigned to an article during the human indexing process, since one of the indexers' tasks is to identify the main topic of the article. We randomly selected fifty abstracts with disorders indexed as the main topic from the abstracts retrieved using PubMed on the five clinical questions described in (Sneiderman *et al.* 2005).

We applied our problem extractor on different segments of the abstract: the title only, the title and first two sentences, and the entire abstract. These results are shown in Table 3. The performance of our best variant (abstract title and first two sentences) approaches the upper bound for MetaMap performance—which is limited by human agreement on the identification of semantic concepts in medical texts, as established in (Pratt & Yetisgen-Yildiz 2003).

Although problem extraction largely depends on disease coverage in UMLS and MetaMap performance, the error rate could be further reduced by more sophisticated recognition of implicitly-stated problems. For example, with respect to a question about immunization in children, an abstract about the measles-mumps-rubella vaccination never mentioned the disease without the word vaccination; hence, no concept of the type DISEASE OR SYNDROME was extracted.

Intervention Extraction

The intervention extractor was evaluated in the same manner as the population extractor and compared to the same baseline. Results are shown in Table 4.

Some of the errors were caused by ambiguity of terms in the intervention. For example, in the clause "serum levels

	correct	unknown	wrong
baseline	60%	-	40%
extractor	80%	-	20%

Table 4: Evaluation of the intervention extractor

of anti-HBsAg and presence of autoantibodies (ANA, ENA) were evaluated", "serum" is recognized as a TISSUE, levels as INTELLECTUAL PRODUCT, and autoantibodies and ANA as IMMUNOLOGIC FACTORS. In this case, however, autoantibodies should be considered a LABORATORY OR TEST RESULT.[2] In other cases, the extraction errors were caused by summary sentences that were very similar to intervention statements, e.g., "This study compared the effects of 52 weeks' treatment with pioglitazone, a thiazolidinedione that reduces insulin resistance, and glibenclamide, on insulin sensitivity, glycaemic control, and lipids in patients with Type 2 diabetes". For this particular abstract, the correct intervention is contained in the following sentence: "Patients with Type 2 diabetes were randomized to receive either pioglitazone (initially 30 mg QD, n = 91) or micronized glibenclamide (initially 1.75 mg QD, n = 109) as monotherapy".

Outcome Extraction

Since outcome statements were annotated in each of the 633 citations in our collection, it was possible to evaluate our outcome extractor on a broader set of abstracts. One hundred and fifty-three citations pertaining to therapy were selected from those not used in the training of the outcome classifiers. Of these, 143 contained outcome statements and were used as the blind held-out test set.

The output of our outcome extractor is a ranked list of sentences. Based on the observation that annotators typically marked two to three sentences in each abstract as outcomes, we evaluated the performance of our extractor at cutoffs of two and three sentences; these results are shown in Table 5, where extractor2 and extractor3 represent the two- and three-sentence cutoffs, respectively. In the evaluation, our outcome extractor was considered correct if the sentences it returned intersected with sentences judged as outcomes by our annotators. Although this is somewhat of a lenient evaluation criteria, we justify it by noting the importance of pointing the physician in the right direction, even if the results are only partially relevant. Motivated by the general expectation that outcome statements are typically found in the conclusion of a structured abstract and near the end of the abstract in the case of unstructured abstracts, we compared our answer extractor to the baseline of returning either the final two or final three sentences in the abstract (base2 and base3 respectively in Table 5).

As can be seen in Table 5, returning the two highest ranked outcome sentences does not outperform either of the baselines. However, we are encouraged by the performance of the outcome extractor at the three-sentence cutoff, where

[2]MetaMap does provide alternative mappings, but the current extraction module only considers the best candidate.

	base2	extractor2	base3	extractor3
correct	74%	75%	75%	95%
unknown	-	-	-	-
wrong	26%	25%	25%	5%

Table 5: Evaluation of the outcome extractor

it achieved higher accuracy than the baselines. The majority of errors in outcome extraction were related to inaccurate sentence boundary identification, chunking errors, and word sense ambiguity in the Metathesaurus.

Sample Output

A complete example of our knowledge extractors working in unison is shown in Figure 1, which presents the extracted PICO elements of the abstract retrieved to answer the following question: "In children with an acute febrile illness, what is the efficacy of single-medication therapy with acetaminophen or ibuprofen in reducing fever?" (Kauffman, Sawyer, & Scheinbaum 1992). "Febrile illness" is the only concept mapped to DISORDER, and hence is identified as the problem. "37 otherwise healthy children aged 2 to 12 years" is correctly identified as the population. "Acetaminophen", "ibuprofen", and "placebo" are correctly extracted as the interventions under study. The three outcome sentences are correctly classified; the short sentence concerning adverse effects was ranked lower than the other three sentences and hence below the cutoff. The study design, from metadata associated with the citation, allows a system to automatically classify this article as a potential level-A answer.

Related Work and Discussion

Clinical question answering is an emerging area of research that has only recently begun to receive serious attention. As a result, there exist relatively few points of comparison to our own work, as the research space is sparsely populated. In this section, however, we will attempt to draw connections to other clinical information systems (although not necessarily for question answering) and related domain-specific question answering systems.

The feasibility of automatically identifying outcome statements in secondary sources has been demonstrated in Niu and Hirst (2004). Their study also illustrates the importance of semantic classes and relations, and in addition suggests an extension of the clinical scenario view as a promising direction in clinical question answering. However, extraction of outcome statements from secondary sources (meta-analyses, in this case) is an easier problem than extraction of outcomes from general MEDLINE citations because secondary sources represent knowledge that has already been distilled by humans (which also limits their scope). Since secondary sources are often more consistently organized, it is possible to depend on certain surface cues for reliable extraction (which is not possible for all MEDLINE abstracts in general). Our study tackles outcome identification in primary medical sources and demonstrates

Antipyretic efficacy of ibuprofen vs acetaminophen
OBJECTIVE–To compare the antipyretic efficacy of ibuprofen, placebo, and acetaminophen. DESIGN–Double-dummy, double-blind, randomized, placebo-controlled trial. SETTING–Emergency department and inpatient units of a large, metropolitan, university-based, children's hospital in Michigan. PARTICIPANTS–*37 otherwise healthy children aged 2 to 12 years*$_{\text{Population}}$ with *acute, intercurrent, febrile illness*$_{\text{Problem}}$. INTERVENTIONS–Each child was randomly assigned to receive a single dose of *acetaminophen*$_{\text{Intervention}}$ (10 mg/kg), *ibuprofen*$_{\text{Intervention}}$ (10 mg/kg) (7.5 or 10 mg/kg), or *placebo*$_{\text{Intervention}}$ (10 mg/kg). MEASUREMENTS/MAIN RESULTS–Oral temperature was measured before dosing, 30 minutes after dosing, and hourly thereafter for 8 hours after the dose. Patients were monitored for adverse effects during the study and 24 hours after administration of the assigned drug. *All three active treatments produced significant antipyresis compared with placebo.*$_{\text{Outcome}}$ *Ibuprofen provided greater temperature decrement and longer duration of antipyresis than acetaminophen when the two drugs were administered in approximately equal doses.*$_{\text{Outcome}}$ No adverse effects were observed in any treatment group. CONCLUSION–*Ibuprofen is a potent antipyretic agent and is a safe alternative for the selected febrile child who may benefit from antipyretic medication but who either cannot take or does not achieve satisfactory antipyresis with acetaminophen.*$_{\text{Outcome}}$

Figure 1: Sample output from our PICO extractors.

that respectable performance is possible with a feature-combination approach.

The literature also contains studies on sentence-level classification of MEDLINE abstracts for other purposes. For example, McKnight and Srinivasan (2003) describe a machine learning approach to automatically label sentences as belonging to introduction, methods, results, or conclusion using structured abstracts as training examples. Note, however, that such labels are orthogonal to PICO frame elements, and hence are not directly relevant to knowledge extraction for clinical question answering. In a similar vein, Light *et al.* (2004) reports on the identification of speculative statements in MEDLINE abstracts.

Other researchers have developed systems that attempt to codify the evidence-based medicine domain model. For example, Cimino and Mendonça studied MeSH terms that are associated with the four basic clinical tasks: etiology, prognosis, diagnosis, and therapy based on analysis 4,000 MEDLINE citations (Mendonça & Cimino 2001). The goal is to automatically classify citations for task-specific retrieval, similar in spirit to the Hedges Project (Wilczynski, McKibbon, & Haynes 2001). The study reported good performance for etiology, diagnosis, and in particular therapy, but not prognosis.

Summarization offers another general approach to building clinical information systems. The PERSIVAL system leverages patient records to generate personalized summaries in response to physicians' queries (McKeown, Elhadad, & Hatzivassiloglou 2003). If patient information is available, deep semantic processing becomes less important, as PERSIVAL is able to achieve respectable performance with relatively superficial techniques. Although patient information is no doubt important to answering clinical questions, information systems that have access to patient records are not widely available. In addition, there are policy concerns and obstacles for such tight integration in a real-world clinical setting.

Our preliminary results have demonstrated the usefulness of knowledge sources in support of question answering. By leveraging existing domain models (in UMLS), software (MetaMap and SemRep), and a task model (PICO frame), semantic knowledge extraction is relatively straightforward, as evidenced by the respectable performance of our population, problem, and intervention extractor using only simple rules. Identification of entities at the conceptual level (i.e., with respect to a semantic class) simplifies extraction of many elements because there is relatively little ambiguity at the semantic level. Successful identification of outcome statements requires a combination of superficial and semantic features, but our results demonstrate the feasibility of this general task. More research is certainly necessary, both to improve performance and develop a more-principled approach to the problem, but we are encouraged by these preliminary results.

The application of domain models and deep semantic knowledge to question answering has been explored by a variety of researchers, e.g., (Jacquemart & Zweigenbaum 2003; Rinaldi *et al.* 2004), and was also a focus at a recent workshop on question answering in restricted domains at ACL 2004. Our work contributes to this ongoing discourse by offering a specific case study in the clinical domain.

Finally, the evaluation of domain-specific question answering systems remains an open research problem. With respect to this issue, Diekema *et al.* (2004) offers interesting observations. It is clear that measures designed for open-domain tasks are not appropriate for the evaluation of systems that only operate on specific domains, but the community has not agreed on a methodology that will allow meaningful comparisons of results from related systems. However, we believe it might be useful to take cues from advances in the evaluation of multi-document summarization (Nenkova & Passonneau 2004) and definition question answering (Lin & Demner-Fushman 2005).

Conclusion

This paper describes knowledge extraction modules that serve as building blocks for a clinical question answering system. Our work is framed within the broader issue of knowledge resources in domain-specific question answering, and how one might leverage domain models. The

preliminary results presented here offer a case study: the recognition of semantic concepts and relations, facilitated by UMLS, MetaMap, and SemRep, simplify the task of knowledge extraction. We are encouraged by these preliminary results, which demonstrate the feasibility of operationalizing major aspects of evidence-based medicine. Information systems in support of the clinical decision-making process have potentially immense impact in affecting the quality of patient care.

Acknowledgements

We would like to thank Barbara Few, Susan Hauser, and Malinda Peeples for their participation in the development of the test collection. The first author is supported by an appointment to the National Library of Medicine Research Participation Program administered by the Oak Ridge Institute for Science and Education through an inter-agency agreement between the U.S. Department of Energy and the National Library of Medicine. The second author would like to thank Kiri for her kind support.

References

Ad Hoc Working Group for Critical Appraisal of the Medical Literature. 1987. A proposal for more informative abstracts of clinical articles. *Annals of Internal Medicine* 106:595–604.

Aronson, A. R. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. In *Proceeding of the American Medical Informatics Association Annual Symposium*, 17–21.

Chambliss, M. L., and Conley, J. 1996. Answering clinical questions. *The Journal of Family Practice* 43:140–144.

Covell, D. G.; Uman, G. C.; and Manning, P. R. 1985. Information needs in office practice: Are they being met? *Annals of Internal Medicine* 103(4):596–599.

DeGroote, S. L., and Dorsch, J. L. 2003. Measuring use patterns of online journals and databases. *Journal of the Medical Library Association* 91(2):231–240.

Demner-Fushman, D.; Hauser, S. E.; Ford, G.; and Thoma, G. R. 2004. Organizing literature information for clinical decision support. In *Proceedings of 11th World Congress on Medical Informatics (MEDINFO 2004)*, 602–606.

Diekema, A. R.; Yilmazel, O.; and Liddy, E. D. 2004. Evaluation of restricted domain question-answering systems. In *Proceedings of the ACL 2004 Workshop on Question Answering in Restricted Domains*.

Ebell, M. H.; Siwek, J.; Weiss, B. D.; Woolf, S. H.; Susman, J.; Ewigman, B.; and Bowman, M. 2004. Strength of Recommendation Taxonomy (SORT): A patient-centered approach to grading evidence in the medical literature. *The Journal of the American Board of Family Practice* 17(1):59–67.

Ely, J. W.; Osheroff, J. A.; Chambliss, M. L.; Ebell, M. H.; and Rosenbaum, M. E. 2005. Answering physicians' clinical questions: Obstacles and potential solutions. *Journal of the American Medical Informatics Association* 12(2):217–224.

Gorman, P. N.; Ash, J. S.; and Wykoff, L. W. 1994. Can primary care physicians' questions be answered using the medical journal literature? *Bulletin of the Medical Library Association* 82(2):140–146.

Jacquemart, P., and Zweigenbaum, P. 2003. Towards a medical question-answering system: A feasibility study. In Baud, R.; Fieschi, M.; Beux, P. L.; and Ruch, P., eds., *The New Navigators: From Professionals to Patients*, volume 95 of *Actes Medical Informatics Europe, Studies in Health Technology and Informatics*. Amsterdam: IOS Press. 463–468.

Kauffman, R. E.; Sawyer, L. A.; and Scheinbaum, M. L. 1992. Antipyretic efficacy of ibuprofen vs acetaminophen. *American Journal of Diseases of Children* 146(5):622–625.

Light, M.; Qiu, X. Y.; and Srinivasan, P. 2004. The language of bioscience: Facts, speculations, and statements in between. In *BioLINK 2004: Linking Biological Literature, Ontologies, and Databases.*, 17–24.

Lin, J., and Demner-Fushman, D. 2005. Automatically evaluating answers to definition questions. Technical Report LAMP-TR-118/CS-TR-4693/UMIACS-TR-2005-03, University of Maryland, College Park.

Lindberg, D. A.; Humphreys, B. L.; and McCray, A. T. 1993. The Unified Medical Language System. *Methods of Information in Medicine* 32(4):281–291.

McCray, A. T.; Burgun, A.; and Bodenreider, O. 2001. Aggregating UMLS semantic types for reducing conceptual complexity. In *Proceedings of 10th World Congress on Medical Informatics (MEDINFO 2001)*, 216–220.

McKeown, K.; Elhadad, N.; and Hatzivassiloglou, V. 2003. Leveraging a common representation for personalized search and summarization in a medical digital library. In *3rd ACM/IEEE 2003 Joint Conference on Digital Libraries*.

McKnight, L., and Srinivasan, P. 2003. Categorization of sentence types in medical abstracts. In *Proceeding of the American Medical Informatics Association Annual Symposium*, 440–444.

Mendonça, E. A., and Cimino, J. J. 2001. Building a knowledge base to support a digital library. In *Proceedings of 10th World Congress on Medical Informatics (MEDINFO 2001)*, 222–225.

Mladenic, D., and Grobelnik, M. 1999. Feature selection for unbalanced class distribution and Naïve Bayes. In *Proceedings of the Sixteenth International Conference on Machine Learning (ICML 1999)*, 258–267.

Nenkova, A., and Passonneau, R. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the 2004 Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT/NAACL 2004)*.

Niu, Y., and Hirst, G. 2004. Analysis of semantic classes in medical text for question answering. In *Proceedings of the ACL 2004 Workshop on Question Answering in Restricted Domains*.

Pratt, W., and Yetisgen-Yildiz, M. 2003. A study of biomedical concept identification: MetaMap vs. people. In *Proceeding of the American Medical Informatics Association Annual Symposium*, 529–533.

Richardson, W. S.; Wilson, M. C.; Nishikawa, J.; and Hayward, R. S. 1995. The well-built clinical question: A key to evidence-based decisions. *American College of Physicians Journal Club* 123(3):A12–A13.

Rinaldi, F.; Dowdall, J.; Schneider, G.; and Persidis, A. 2004. Answering questions in the genomics domain. In *Proceedings of the ACL 2004 Workshop on Question Answering in Restricted Domains*.

Rindflesch, T. C., and Fiszman, M. 2003. The interaction of domain knowledge and linguistic structure in natural language processing: Interpreting hypernymic propositions in biomedical text. *Journal of Biomedical Informatics* 36(6):462–477.

Sackett, D. L.; Strauss, S. E.; Richardson, W. S.; Rosenberg, W.; and Haynes, R. B. 2000. *Evidence-Based Medicine: How to Practice and Teach EBM*. Churchill Livingstone, second edition.

Sneiderman, C.; Demner-Fushman, D.; Fiszman, M.; and Rindflesch, T. C. 2005. Semantic characteristics of MEDLINE citations useful for therapeutic decision-making. In *Proceeding of the American Medical Informatics Association Annual Symposium*. Under review.

Wilczynski, N.; McKibbon, K. A.; and Haynes, R. B. 2001. Enhancing retrieval of best evidence for health care from bibliographic databases: Calibration of the hand search of the literature. In *Proceedings of 10th World Congress on Medical Informatics (MEDINFO 2001)*, 390–393.

Yang, Y., and Pedersen, J. O. 1997. A comparative study on feature selection in text categorization. In *Proceedings of the 14th International Conference on Machine Learning*, 412–420.